Appendix

A Theoretical Proof

We provide a detailed derivation and proof of the least-squares pose regression process in the text. The camera position parameters θ and $\boldsymbol{t}$ can be obtained by solving the following loss function.

$$\xi = \arg\min_{\theta,\boldsymbol{t}} \sum_{i=1}^{n} S_i(\boldsymbol{R}\boldsymbol{p}'_i + \boldsymbol{t} - \hat{\boldsymbol{p}}_i)^2, \boldsymbol{R} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \tag{8}$$

The relationship between the corresponding points $\boldsymbol{p}'_i$ and $\hat{\boldsymbol{p}}_i$ in point sets $\boldsymbol{p}'$ and $\hat{\boldsymbol{p}}$ can be expressed as $\boldsymbol{p}'_i = \boldsymbol{R}\hat{\boldsymbol{p}}_i + \boldsymbol{t}$. To simplify the problem, we calculate the centroids $\boldsymbol{g'}$ and $\hat{\boldsymbol{g}}$ of point sets $\boldsymbol{p}'$ and $\hat{\boldsymbol{p}}$, respectively. We then subtract the centroids from each point in their respective sets, resulting in $\boldsymbol{q'}_i = \boldsymbol{p}'_i - \boldsymbol{g'}$ and $\hat{\boldsymbol{q}}_i = \hat{\boldsymbol{p}}_i - \hat{\boldsymbol{g}}$. This simplifies the problem to:

$$\begin{aligned} \xi &= \arg\min_{\theta} \sum_{i=1}^{n} S_i(\boldsymbol{R}\boldsymbol{q}'_i - \hat{\boldsymbol{q}}_i)^2 \\ &= \arg\min_{\theta} \sum_{i=1}^{n} S_i(\boldsymbol{R}\boldsymbol{q}'_i - \hat{\boldsymbol{q}}_i)^t(\boldsymbol{R}\boldsymbol{q}'_i - \hat{\boldsymbol{q}}_i) \\ &= \arg\min_{\theta} \sum_{i=1}^{n} S_i({\boldsymbol{q}'_i}^t\boldsymbol{q}'_i) + {\hat{\boldsymbol{q}}_i}^t\hat{\boldsymbol{q}}_i - 2{\hat{\boldsymbol{q}}_i}^t\boldsymbol{R}\boldsymbol{q}'_i \end{aligned} \tag{9}$$

Therefore, minimizing ξ is equivalent to maximizing $\sum_{i=1}^{n} S_i({\hat{\boldsymbol{q}}_i}^t\boldsymbol{R}\boldsymbol{q}'_i)$. Assuming that $\boldsymbol{H} = \sum_{i=1}^{n} S_i(\boldsymbol{q}'_i{\hat{\boldsymbol{q}}_i}^t)$. then

$$\sum_{i=1}^{n} S_i({\hat{\boldsymbol{q}}_i}^t\boldsymbol{R}\boldsymbol{q}'_i) = tr(\sum_{i=1}^{n} \boldsymbol{R}S_i\boldsymbol{q}'_i{\hat{\boldsymbol{q}}_i}^t) = tr(\boldsymbol{R}\boldsymbol{H}) \tag{10}$$

Upon performing an SVD decomposition on $\boldsymbol{H}$, the result is $\boldsymbol{H} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^t$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices of dimensions 3×3, and $\boldsymbol{\Lambda}$ is 3×3 diagonal matrix with non-negative elements.

Then we introduce an orthogonal matrix $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^t$, and observe that $\boldsymbol{X}\boldsymbol{H} = \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^t$ is a symmetric positive-definite matrix. According to the properties of positive definite matrices, it holds that $tr(\boldsymbol{X}\boldsymbol{H}) \geq tr(\boldsymbol{B}\boldsymbol{X}\boldsymbol{H})$ for any 3×3 orthogonal matrix $\boldsymbol{B}$. Therefore, we obtain the maximum value of $\sum_{i=1}^{n} S_i({\hat{\boldsymbol{q}}_i}^t\boldsymbol{R}\boldsymbol{q}'_i)$ when $\boldsymbol{R} = \boldsymbol{X}$, which implies that

$$\boldsymbol{R} = \boldsymbol{U}\boldsymbol{V}^t \tag{11}$$

The dense matching relationship between the graphs gives rise to the uniqueness of solution for R. Subsequently, the relative displacement can be determined as

$$\boldsymbol{t} = \boldsymbol{g}' - \boldsymbol{R}\hat{\boldsymbol{g}} \tag{12}$$

By utilizing equations 11 and 12, we can obtain the camera location and azimuth through a straightforward differentiable process, using the least-squares regression method.

B Localization qualitative results

In Figures 5 and 6, we present additional qualitative cross-view localization results on the KITTI, Ford multi-AV, VIGOR, and Oxford RobotCar datasets.

As pointed out in the main paper Section 4.6 of the main text, our method tends to exhibit poorer performance in terms of longitudinal localization accuracy. We present some examples of localization errors in Figure 5. In the visualization results it can be noticed that our algorithm demonstrates higher accuracy in lateral localization along the direction of travel, while performing poorly in longitudinal localization, particularly in the Ford multi-AV dataset of suburban scenes. The displayed examples of localization failures reveal that the limited horizontal field of view and the absence of reference objects along the sides of the lane contribute to the localization errors. In such scenarios, our algorithm is still able to rely on lane features to accomplish lateral localization compared to LM. However, neither method performs well in longitudinal localization.

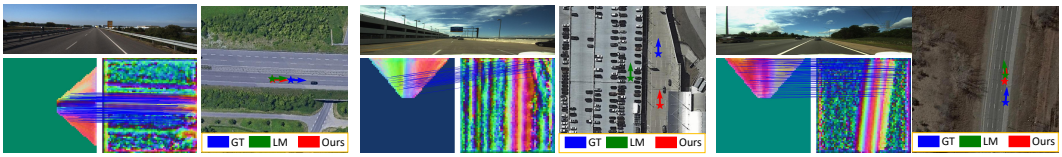
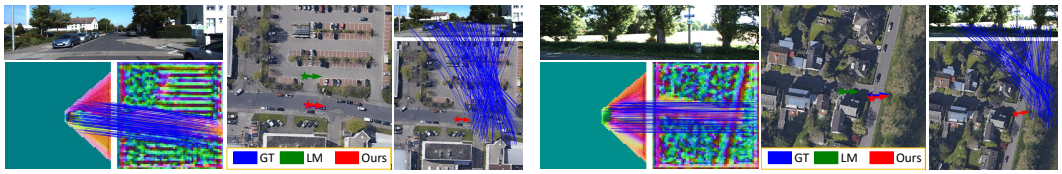
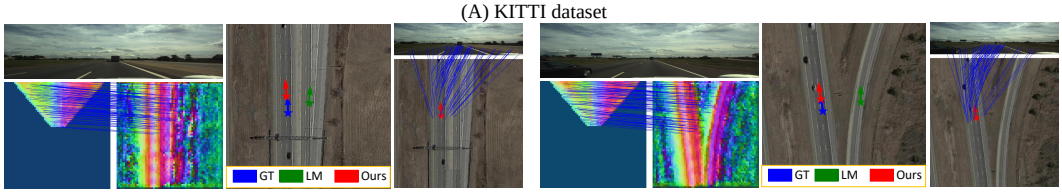


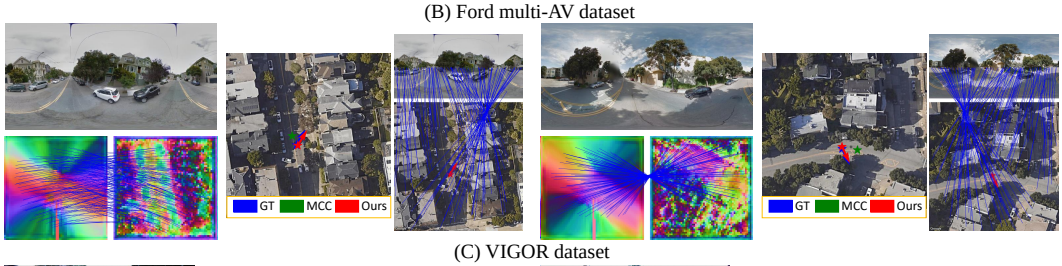
Figure 5: Failure cases. The left image sourced from KITTI dataset and the two on the right from the Ford multi-AV dataset.For each scene, the up left is the ground image, the bottom left denotes matching inliers and the right shows the satellite image and localization results. For visual simplicity, dense matching displays only partial inliers.



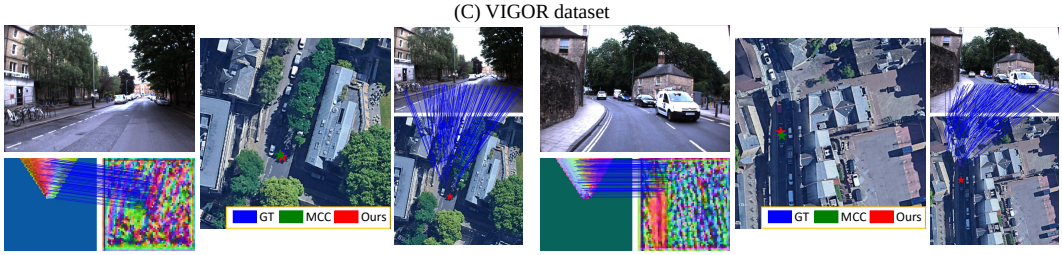
(A) KITTI dataset



(B) Ford multi-AV dataset



(C) VIGOR dataset



(D) Oxford RobotCar dataset

Figure 6: Qualitative results on the four datasets. For each scene, the up left is the ground image, the bottom left denotes matching inliers, the center shows the satellite image and localization results, and the right displays a visualization of the matched points mapped back to the original image. The central axis of a panoramic image represents its orientation. For visual simplicity, dense matching displays only partial inliers.

In addition, we have provided more visualizations for all four datasets. As discussed in Section 4 of the main text, our approach outperforms previous methods. It is important to note that, during testing on the VIGOR dataset, the MCC method cannot handle angle noise. Therefore, the visualization for MCC is based on a scenario where the angle noise is zero, while our visualization results are obtained in a scenario where the angle is completely unknown. Furthermore, to demonstrate the effectiveness of dense matching, we visualize the dense matching points mapped onto the original images in Figure 5. It can be observed that our pipeline establishes correspondences between satellite and ground-level image points, achieving precise localization of ground cameras in satellite images. This further confirms the efficacy of the architecture proposed in our work for cross-view localization using dense pixel flow fields.